Marie Sbrocca
11/7/16
Paper #4 -- Proposing an Experiment

Thus far, my research topic has addressed human interactions with conversational AI interfaces. Specifically, I have proposed to investigate user reported satisfaction with these interactions and changes in use patterns over time. Due to the differences in individual interface designs and their relative affordances, I have specifically addressed interaction data for Amazon's Alexa (Echo) interface.

For the purpose of this proposal, I will seek to answer the following question using a randomized control trial experiment: Does affective priming[1] have an effect on user satisfaction with interactions with conversational AIs? Because Alexa's interface is bimodal (mobile app and voice), and requires two different interaction modalities, I will also evaluate the effect of modality on priming. To conduct this experiment, I will employ the following research design, again using the Amazon Alexa interface.

Users interact with Alexa via two different modalities: a mobile application, and a voice-based interface. Interactions using the voice-based interface are structured as questions or directions, and interactions using the mobile application are used to collect feedback. In the voice based modality, the user receives a response from Alexa each time they interact with it (i.e., if Alexa understands the user's request, it executes its interpretation of the appropriate action; if it does not, it tells the user that it did not understand). In the app, the user is presented with questions (cues) that allow them to provide feedback about the quality of their interactions.

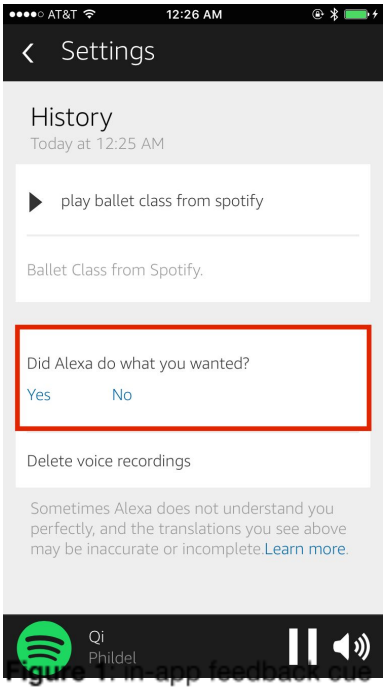

Figure 1: In-app feedback cue

Currently, these responses and cues are neutral. When Alexa does not understand a query, it responds with something along the lines of "Hmm. I can't find the answer to the question I heard." In the app, the user is simply presented with the option to indicate if Alexa did what they wanted for a given interaction: Did Alexa do what you wanted? (Figure 1).

To study the effects of affective priming, I will prepare responses and cues that vary in their positivity or negativity, such that there are sets of neutral, positive, and negative responses to unclear queries to Alexa, and neutral, positive, and negative cues for feedback in the app.

Subjects will then be divided into nine groups. The control group will receive the default, neutral responses and cues from Alexa in-app and through the voice interface. The other eight groups will be given combinations of neutral, positive, and negative affective responses and cues as defined in Table 1.

Group	Voice modality (response)	App modality (cue)

[1] Affective priming occurs when emotionally salient content affects subsequent actions or attitudes.

1 (control)	neutral	neutral
2 (positive)	positive	positive
3 (negative)	negative	negative
4 (modality control, app positive)	neutral	positive
5 (modality control, app negative)	neutral	negative
6 (modality control, voice positive)	positive	neutral
7 (modality control, voice negative)	negative	neutral
8 (mixed 1, app negative)	positive	negative
9 (mixed 2, voice negative)	negative	positive

Table 1: experiment groups

This is an example of a fully digital experiment: participant recruitment, randomization, treatments delivery, and outcome measurement are all digital. Subjects would be drawn from the existing user base of this interface. Randomization would occur by randomly assigning treatments to different users who are already stored in a database. Treatments would then be delivered to these users via modifications to the content they receive via a digital interface, and outcomes would be measured as part of interaction data collection (i.e. yes/no responses, use patterns over time relative to treatment groups).

This experiment will be conducted in the field, but reaps some of the benefits of lab research: because of the nature of subjects (existing users of a digital interface), pre-treatment information, such as demographic information is available. This experiment design is also beneficial because its cost do not increase proportionally to scale.

This experiment is a randomized control trial, and will employ a “between subjects” design. However, it may also be feasible to shift this experiment to a mixed design and also evaluate a “within subjects” design. Because user interaction data is continuously collected, there may be data available from users in non-control groups from prior to this experiment that meets control group criteria (contingent upon other changes, updates, or A/B testing that have occurred within this interface).

Internal validity for this project will be high: because this is a fully digital experiment, it will be straightforward to administer treatments and measure results as intended. Statistical validity should also be high, though caution must be taken with regard to randomly sampling this group in order to make sure no groups are over or under-represented. Additionally, because of the pre-existing concept that humans weight negative feedback more heavily than positive feedback, it may be necessary to weight the value of responses accordingly (Baumeister et al.). Construct validity should be high: I am appropriately applying existing concepts about priming, perception, and multimodal interactions from cognitive psychology and HCI research. While multiple interaction modalities make confounds more likely, I have addressed this by employing a factorial research design. External validity may present confounds: multimodal interfaces are still relatively uncommon, and there is not a large body of literature to draw from to validate this. Additionally, since this experiment occurs in the field, there are a number of other possible environment confounds for each interaction (that is, what else is going on around the user). While outside the scope of this proposal, additional research could be conducted on other multimodal interfaces to help address external validity.

It will be very important to ensure that heterogenous treatment effects are appropriately represented, and that no sub-group of users is over-represented in this study. In this case, the amount that users interact with the interface seems to be the most likely confound with regard to heterogeneity. The amount the users interact with a given product varies substantially across the population, and in a purely randomized, population level experiment, the individuals who use the product more often may be over-represented. I plan to address this by using pre-treatment data on individual user patterns to identify potential confounds.

I have utilized factorial experiment design in order to identify causal mechanisms in this experiment. Employing this design will allow me to explore causality by individually comparing each possible combination of affect and modality. I will be able to segment users into wholly neutral (control), positive, and negative groups. I will also see users who received a positive or negative for one modality and neutral for the other (which should help identify modality-specific causality), as well as four mixed groups, which should help identify strength of causality for modality.

This experiment design would require substantial buy-in from a corporate partner, in this case, Amazon's Alexa team. I recognize that this may present its own set of issues, or render this research design unfeasible within the context of academic research. However, to conduct interaction research that evaluates priming effects from content inherent to an interface, it would be essential to work with a group that can actually modify a product that is used in natural environments.

Because this experiment evaluates affective priming, I suspect it would run into some of the same criticisms of the Facebook emotional contagion study. However, there are precedents in HCI research on affective priming in digital contexts that include emotionally negative stimuli that have not been subjected to similar fallout. I postulate that since this study alters emotions in such a way that is directed toward a digital agent (i.e., Alexa), not general mood (i.e., Facebook contagion), it would be less likely to be the object of such concern.

Works Cited

Baumeister, Roy F., Ellen Bratlavsky, Catrin Finkenauer, and Kathleen D. Vohs. "Bad Is Stronger Than Good." Review of General Psychology, 2001. Web. http://dare.ubvu.vu.nl/bitstream/handle/1871/17432/Baumeister_Review?sequence=2.

Salganik, Matthew J. "Bit By Bit: Social Research in the Digital Age." Bit By Bit: Social Research in the Digital Age. Accessed October 24, 2016. http://www.bitbybitbook.com/.